

ON TESTING INDEPENDENCE AND GOODNESS-OF-FIT IN LINEAR MODELS

Arnab Sen and Bodhisattva Sen*

University of Minnesota and Columbia University

Abstract

We consider a linear regression model and propose an omnibus test to simultaneously check the assumption of independence between the error and the predictor variables, and the goodness-of-fit of the parametric model. Our approach is based on testing for independence between the residual and the predictor using the recently developed Hilbert-Schmidt independence criterion, see [GFT⁺08]. The proposed method requires no user-defined regularization, is simple to compute, based merely on pairwise distances between points in the sample, and is consistent against all alternatives. We develop the distribution theory of the proposed test-statistic, both under the null and the alternative hypotheses, and devise a bootstrap scheme to approximate its null distribution. We prove the consistency of the bootstrap procedure. The superior finite sample performance of our procedure is illustrated through a simulation study.

Keywords: Bootstrap, goodness-of-fit test, linear regression, model checking, reproducing kernel Hilbert space, test of independence

1 Introduction

In regression analysis, given a random vector (X, Y) where X is a d_0 -dimensional predictor and Y is the one-dimensional response, we want to study the relationship

*Supported by NSF Grants DMS-1150435 and AST-1107373

between Y and X . The relationship can always be summarized as:

$$Y = m(X) + \eta, \quad (1)$$

where m is the regression function and $\eta := Y - m(X)$ is the error that has conditional mean 0 (given X). However, in practice, we usually assume that $X \perp\!\!\!\perp \eta$, i.e., η is independent of X . Moreover, in linear regression, we assume that m belongs to a parametric class, e.g.,

$$\mathcal{M}_\beta := \{g(x)^\top \beta : \beta \in \mathbb{R}^d\}, \quad (2)$$

where $g(x) = (g_1(x), \dots, g_d(x))^\top$ is the set of known (measurable) predictor functions, and β is the finite-dimensional unknown (coefficient) parameter.

In this paper we propose an omnibus test to check simultaneously the assumption of independence between X and η , and the goodness-of-fit of the linear regression model, i.e., test the null hypothesis

$$H_0 : X \perp\!\!\!\perp \eta, m \in \mathcal{M}_\beta, \quad (3)$$

given i.i.d. data $(X_1, Y_1), \dots, (X_n, Y_n)$ from the regression model (1). Even when we consider the predictor variables fixed, i.e., we condition on X , our procedure can be used to check whether the conditional distribution of η given X depends on X . This will, in particular, help us detect heteroscedasticity (i.e., whether the conditional variance of η given X depends on X) and departures from the assumption of homoscedasticity of the errors. As far as we are aware, there is no test that can simultaneously check for these two crucial model assumptions in linear regression.

Our procedure is based on testing for the independence of X and the residual $\hat{\eta}$ (obtained from fitting the parametric model) using the recently developed Hilbert-Schmidt independence criterion (HSIC); see [GFT⁺08]. Among the virtues of this test is that it is automated (i.e., requires no user-defined regularization), extremely simple to compute, based merely on the distances between points in the sample, and is consistent against all alternatives. Also, compared to other measures of dependence, HSIC does not require any smoothness assumption on the joint distribution of X and η (e.g., existence of a density), and its implementation is not computationally intensive for higher dimensions (i.e., when d_0 is large). Moreover, this independence testing procedure also yields a novel approach to testing for the goodness-of-fit of the fitted regression model: under model mis-specification, the residuals, although uncorrelated with the predictors (by definition of the least squares procedure), are very much dependent on the predictors, and the HSIC can detect this dependence; under H_0 , the test statistic exhibits n^{-1} -rate of convergence, whereas, under dependence, we observe $n^{-1/2}$ -rate of convergence for the centered test statistics.

We find the limiting distribution of the test statistic, under both the null and alternative hypotheses. Interestingly, we see that the asymptotic distribution is very different from that if the true error η were observed. To approximate the null distribution of the test statistic, we propose a bootstrap scheme and prove its consistency. Note that the usual permutation test, which is used quite often in testing independence, cannot be directly used in this scenario as we do not observe η .

Over the last two decades several tests for the goodness-of-fit of a parametric model have been proposed under varied conditions on the distribution of the errors and its dependence on the predictors. [CKWY88] introduced tests of the null hypothesis that a regression function has a particular parametric structure under the assumption of independent homoscedastic normal errors. [AB93] used nonparametric regression to check linear relationships under independent homoscedastic errors; also see [ES90] and [HM93]. In [FH01], some tests are proposed for examining the adequacy of a family of parametric models against large nonparametric alternatives under the assumption of independence and normality of the errors. In [GL05] the authors propose data-driven smooth tests for a parametric regression function. For other relevant work on this topic see [CS10], [Xia09] and the references therein. Any test using a nonparametric regression estimator runs into an ill-posed problem requiring the delicate choice of a number of tuning parameters, e.g., smoothing parameter(s). However, a few alternative approaches have been developed that circumvent these problems; see e.g. [Bie90], [Stu97]. However, most of these tests are difficult to implement and do not usually work well when the dimension of X is not low.

Although the independence of error and covariate is a common assumption in regression, there are very few methods available in the literature to test this. However, there has been work on testing for heteroscedasticity between the error and the predictor; see e.g., [CW83], [BP79], [Ken08] and the references therein. In the nonparametric setup, [EVK08b] and [EVK08a] propose tests for independence but only for univariate covariates. Generalization to the multivariate case is recently considered in [NVK10]; also see [Neu09].

The paper is organized as follows: in Section 2 we introduce the HSIC and discuss other measures of dependence. We formulate the problem and state our main results in Section 3. A bootstrap procedure to approximate the distribution of the test statistic is developed in Section 4. A finite sample study of our method along with some well-known competing procedures is presented in Section 5. In Section 6 we present a result on triangular arrays of random variables that will

help us understand the limiting behavior of our test statistic under the null and alternative hypotheses, and yield the consistency of our bootstrap approach. The proofs of the main results are given in Section 7.

2 Testing independence of two random vectors

In this section we briefly review the HSIC for testing the independence of two random vectors; for a more systematic study of the procedure see [GFT⁺08], [GBSS05], [SSGF12]. We start with some background and notation. By a reproducing kernel Hilbert space (RKHS) \mathcal{F} of functions on a domain \mathcal{U} with a positive definite kernel $k : \mathcal{U} \times \mathcal{U} \rightarrow \mathbb{R}$ we mean a Hilbert space of functions from \mathcal{U} to \mathbb{R} with inner product $\langle \cdot, \cdot \rangle$, satisfying the reproducing property:

$$\langle f(u), k(u, \cdot) \rangle = f(u),$$

for all $f \in \mathcal{F}$ and $u \in \mathcal{U}$. We say that \mathcal{F} is *characteristic* if and only if the map

$$P \mapsto \int_{\mathcal{U}} k(\cdot, u) dP(u),$$

is injective on the space of all Borel probability measures on \mathcal{U} such that $\int_{\mathcal{U}} k(u, u) dP(u) < \infty$. Likewise, let \mathcal{G} be a second RKHS on a domain \mathcal{V} with positive definite kernel l . Let P_{uv} be a Borel probability measure defined on $\mathcal{U} \times \mathcal{V}$, and let P_u and P_v denote the respective marginal distributions on \mathcal{U} and \mathcal{V} . Assuming that

$$\mathbb{E}[k(U, U)] < \infty \quad \text{and} \quad \mathbb{E}[l(V, V)] < \infty, \quad (4)$$

where $(U, V) \sim P_{uv}$, the HSIC of P_{uv} is defined as

$$\begin{aligned} \theta(U, V) &:= \mathbb{E}[k(U, U')l(V, V')] + \mathbb{E}[k(U, U')] \mathbb{E}[l(V, V')] \\ &\quad - 2 \mathbb{E}[k(U, U')l(V, V'')], \end{aligned} \quad (5)$$

where $(U', V'), (U'', V'')$ are i.i.d. copies of (U, V) . It is not hard to see that $\theta(U, V) \geq 0$. More importantly, when \mathcal{F} and \mathcal{G} are *characteristic* (see [Lyo11], [SSGF12]), and (4) holds, then

$$\theta(U, V) = 0 \quad \text{if and only if} \quad P_{uv} = P_u \times P_v.$$

Given an i.i.d. sample $(U_i, V_i)_{1 \leq i \leq n}$ from P_{uv} , we want to test whether P_{uv} factorizes as $P_u \times P_v$. For the purpose of testing independence, we will use a biased

empirical estimate of θ [[GBSS05], Definition 2], obtained by replacing the unbiased U -statistics with the V -statistic

$$\hat{\theta}_n := \frac{1}{n^2} \sum_{i,j}^n k_{ij} l_{ij} + \frac{1}{n^4} \sum_{i,j,q,r}^n k_{ij} l_{qr} - 2 \frac{1}{n^3} \sum_{i,j,q}^n k_{ij} l_{iq} = \frac{1}{n^2} \text{trace}(KHLH), \quad (6)$$

where the summation indices denote all t -tuples drawn with replacement from $\{1, \dots, n\}$, t being the number of indices below the sum, $k_{ij} := k(U_i, U_j)$, and $l_{ij} := l(V_i, V_j)$, K and L are $n \times n$ matrix with entries k_{ij} and l_{ij} , respectively, $H := \mathbf{I} - n^{-1} \mathbf{1} \mathbf{1}^\top$, and $\mathbf{1}$ is an $n \times 1$ vector of ones. Note that the cost of computing this statistic is $O(n^2)$.

Examples of translation invariant characteristic kernel functions on \mathbb{R}^p , for $p \geq 1$, include the Gaussian radial basis function kernel $k(u, u') = \exp(-\sigma^{-2} \|u - u'\|^2)$, $\sigma > 0$, the Laplace kernel $k(u, u') = \exp(-\sigma^{-1} \|u - u'\|)$, the inverse multiquadratics $k(u, u') = (\beta + \|u - u'\|^2)^{-\alpha}$, $\alpha, \beta > 0$, etc. We will use the Gaussian kernel in our simulation results.

One can, in principle, use any other test of independence and develop a theory parallel to ours. The choice of the HSIC is motivated by a number of computational and theoretical advantages, see e.g., [GBSS05] and [GFT⁺08]. It is worth mentioning in this regard that the recently developed method of distance covariance, introduced by [SRB07] and [SR09], has received much attention in the statistical community. It tackles the problem of testing and measuring dependence between two random vectors in terms of a weighted L^2 -distance between characteristic functions of the joint distribution of two random vectors and the product of their marginals; see [SSGF12] for a comparative study of the HSIC and the distance covariance methods. However, the kernel induced by the semi-metric used in the distance covariance method (see [SSGF12]) is not smooth and hence is difficult to study theoretically, at least using our techniques.

3 Our test statistic

We will consider the regression model (1). We denote by $Z = (X, \eta) \sim P$ where $Z \in \mathbb{R}^{d_0} \times \mathbb{R}$ and $\mathbb{E}(\eta|X) = 0$. Let P_X and P_η be the marginal distributions of X and η respectively. To start with, we will assume that m does not necessarily belong to \mathcal{M}_β , as defined in (2). Assuming that $\mathbb{E}[g(X)g(X)^\top] < \infty$, $\mathbb{E}[m(X)^2] < \infty$ and $\mathbb{E}[\eta^2] < \infty$, let us define

$$D^2(\beta) := \mathbb{E}[(Y - g(X)^\top \beta)^2],$$

for $\beta \in \mathbb{R}^d$. From the definition of m , we see that

$$D^2(\beta) := \mathbb{E}[(Y - m(X))^2] + \mathbb{E}[(m(X) - g(X)^\top \beta)^2].$$

The function D^2 is minimized at $\tilde{\beta}_0$ if and only if $\tilde{\beta}_0$ is a minimizer of

$$\tilde{D}^2(\beta) := \mathbb{E}[(m(X) - g(X)^\top \beta)^2].$$

The quantity $\tilde{D}^2(\tilde{\beta}_0)$ measures the distance between the true m and the hypothetical model \mathcal{M}_β . Clearly, if $m(X) = g(X)^\top \beta_0$, then $\beta_0 = \tilde{\beta}_0$. Under the assumption that $\mathbb{E}[g(X)g(X)^\top]$ is invertible, $\tilde{D}^2(\beta)$ has the unique minimizer

$$\tilde{\beta}_0 = \mathbb{E}[g(X)g(X)^\top]^{-1} \mathbb{E}[m(X)g(X)].$$

Thus, $g(x)^\top \tilde{\beta}_0$ is the “closest” function (in the least squares sense) to $m(x)$ in \mathcal{M}_β .

Given i.i.d. data $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ from the regression model (1), we compute the least squares estimator (LSE) in the class \mathcal{M}_β as

$$\hat{\beta}_n := \operatorname{argmin}_{\beta \in \mathbb{R}^d} \sum_{i=1}^n (Y_i - g(X_i)^\top \beta)^2. \quad (7)$$

Letting

$$A_n := \frac{1}{n} \sum_{i=1}^n g(X_i)g(X_i)^\top,$$

note that

$$\hat{\beta}_n = A_n^{-1} \left(\frac{1}{n} \sum_{i=1}^n g(X_i)Y_i \right),$$

provided, of course, that A_n is invertible. Let

$$e_i := Y_i - g(X_i)^\top \hat{\beta}_n, \quad (8)$$

for $i = 1, \dots, n$, be the *observed* residuals. The test statistic we consider is

$$T_n = \frac{1}{n^2} \sum_{i,j} k_{ij} l_{ij} + \frac{1}{n^4} \sum_{i,j,q,r} k_{ij} l_{qr} - 2 \frac{1}{n^3} \sum_{i,j,q} k_{ij} l_{iq}, \quad (9)$$

where $k_{ij} := k(X_i, X_j)$, and $l_{ij} := l(e_i, e_j)$ with k and l being characteristic kernels defined on $\mathbb{R}^{d_0} \times \mathbb{R}^{d_0}$ and $\mathbb{R} \times \mathbb{R}$ respectively. Note that our test statistic is almost identical to the empirical estimate $\hat{\theta}_n$ of HSIC between X and η described in (6) except for the fact that we replace the unobserved errors η_i 's by the observed residuals e_i 's.

For any $u = (u_1, \dots, u_p) \in \mathbb{R}^p$, we define the ℓ_∞ -norm of u as $|u|_\infty = \max_{1 \leq i \leq p} |u_i|$. We will assume throughout the paper that

(I) $A := \mathbb{E}[g(X)g(X)^\top]$ is invertible.

Moreover, we will always assume the following conditions on the kernels k, l .

(K) The kernels k and l are characteristic kernels. k is continuous and l is twice continuously differentiable. Denoting the partial derivatives of l as $l_x(x, y) := \partial_x l(x, y)$, $l_{xy}(x, y) := \partial_x \partial_y l(x, y)$, etc., we assume that l_{xx} , l_{xy} and l_{yy} are Lipschitz continuous with Lipschitz constant L (w.r.t. the ℓ_∞ -norm).

We study the behavior of the test statistic T_n under the null hypothesis (3), and also under the following different scenarios:

$$\begin{aligned} H_1 : & \quad X \not\perp \eta, m \in \mathcal{M}_\beta, \\ H_2 : & \quad X \perp \eta, m \notin \mathcal{M}_\beta, \\ H_3 : & \quad X \not\perp \eta, m \notin \mathcal{M}_\beta. \end{aligned} \tag{10}$$

To find the limiting distribution of T_n under H_0 , we will assume the following set of moment conditions on X and η .

(M) (a) $\mathbb{E}[|g(X)|_\infty^2] < \infty$,

(b) $\mathbb{E}[\eta^2] < \infty$,

(c) for any $1 \leq q, r, s, t \leq 4$,

$$\mathbb{E} [k^2(X_q, X_r)(1 + |g(X_s)|_\infty^2)(1 + |g(X_t)|_\infty^2)] < \infty,$$

(d) for any $1 \leq q, r \leq 2$,

$$\mathbb{E}[f^2(\eta_q, \eta_r)] < \infty \text{ for } f = l, l_x, l_y, l_{xx}, l_{yy}, l_{xy}.$$

Theorem 3.1 *Suppose that conditions (I), (K) and (M) hold. Then, under H_0 ,*

$$nT_n \rightarrow_d \chi,$$

where χ has a non-degenerate distribution which depends upon $P = P_X \times P_\eta$ and is denoted by $\chi = \chi(P_X \times P_\eta)$.

Remark 3.1 *The random variable χ can be expressed as a quadratic function of a Gaussian field with a certain nontrivial covariance structure. This is in contrast with the limiting description of degenerate V-statistics where the limiting random variable can be described as a quadratic function of a family of independent Gaussian random variables. The explicit description of χ is slightly complicated and can be easily recovered from the proof; see (31) in Section A.1.3. However, from a practical point of view, such a description is of little use, since P is unknown to the user.*

Remark 3.2 *Though one might be tempted to believe that replacing the (unobserved) true errors η_i by the residuals e_i should not alter the limiting distribution of the test statistic, it turns out that the two limiting distributions are significantly different; see Figure 1.*

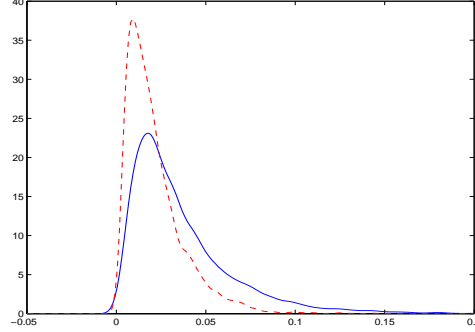


Figure 1: Estimated density of nT_n obtained with the true unknown errors (in solid blue) and the estimated residuals (in dashed red) in the linear model $Y = 1 + X + \eta$, where $\eta \sim N(0, \sigma^2 = 0.1)$, $X \sim N(0, 1)$, $X \perp \eta$, and $n = 100$.

Next we will study the limiting behavior of our test statistic T_n under the different alternatives H_1, H_2 and H_3 as in (10). We first introduce the error under model mis-specification as

$$\epsilon := m(X) - g(X)^\top \tilde{\beta}_0 + \eta.$$

Note that when $m \in \mathcal{M}_\beta$, $\epsilon \equiv \eta$. We assume the following set of moment conditions for H_1, H_2 and H_3 .

$$(M') \text{ (a) } \mathbb{E}[|g(X)|_\infty^2] < \infty \text{ and } \mathbb{E}[m(X)^2] < \infty$$

$$(b) \mathbb{E}[\eta^2] < \infty \text{ and } \mathbb{E}[|g(X)|_\infty^2 \epsilon^2] < \infty$$

$$(c) \text{ for any } 1 \leq q, r, s, t \leq 4,$$

$$(i) \mathbb{E}[k^2(X_q, X_r) l^2(\epsilon_s, \epsilon_t)] < \infty,$$

$$(ii) \mathbb{E}[|k(X_q, X_r)| |\nabla l(\epsilon_s, \epsilon_t)|_\infty (|g(X_s)|_\infty + |g(X_t)|_\infty)] < \infty,$$

$$(iii) \mathbb{E}[|k(X_q, X_r)| (|g(X_s)|_\infty^3 + |g(X_t)|_\infty^3)] < \infty,$$

$$(iv) \mathbb{E}[|k(X_q, X_r)| |\text{Hess}(l)(\epsilon_s, \epsilon_t)|_\infty (|g(X_s)|_\infty^2 + |g(X_t)|_\infty^2)] < \infty,$$

where $|\nabla l(\epsilon_s, \epsilon_t)|_\infty = \max(|l_x(\epsilon_s, \epsilon_t)|, |l_y(\epsilon_s, \epsilon_t)|)$ and

$|\text{Hess}(l)(\epsilon_s, \epsilon_t)|_\infty = \max(|l_{xx}(\epsilon_s, \epsilon_t)|, |l_{xy}(\epsilon_s, \epsilon_t)|, |l_{yy}(\epsilon_s, \epsilon_t)|)$. Here $\epsilon_1, \dots, \epsilon_4$ are i.i.d. copies of ϵ .

Theorem 3.2 *Suppose that conditions (I), (K) and (M') hold. Assume further under H_2 that $m(X) - g(X)^\top \tilde{\beta}_0$ is not a constant. Then*

$$\sqrt{n}(T_n - \theta) \rightarrow_d N(0, \sigma^2),$$

where $\theta = \theta(X, \epsilon)$ is defined in (5) and $\theta > 0$. The variance σ^2 depends on the joint distribution of (X, ϵ) and an expression for it can be found in (18) in the proof of the theorem.

4 Consistency of the bootstrap

Theorem 3.1 is not very useful in computing the cut-off value to test H_0 , using the statistic nT_n , as the asymptotic distribution χ involves infinitely many nuisance parameters. An obvious alternative in such a situation is to use a resampling technique to approximate the cut-off. In independence testing problems, a natural choice is to use the permutation test; see e.g., [SR09], [GFT⁺08].

However, as we are using the residuals e_i 's instead of the true unknown errors η_i 's in our test statistic, a permutation based test will not work. Indeed, under the null hypothesis, the joint distribution of $(X_i, e_{\pi(i)})_{1 \leq i \leq n}$ is not invariant under the permutation π , even though the joint distribution of $(X_i, \eta_{\pi(i)})_{1 \leq i \leq n}$ remains unchanged under π .

In this section we show that the bootstrap can be used to consistently approximate the distribution of nT_n , under H_0 . In the following we describe our bootstrap procedure.

1. Let \mathbb{P}_{n, e° be the empirical distribution of centered residuals, i.e.,

$$e_i^\circ = e_i - \bar{e},$$

for $i = 1, \dots, n$, where e_i 's are defined in (8) and $\bar{e} = \frac{1}{n} \sum_{i=1}^n e_i$. Let $\mathbb{P}_{n, X}$ be the empirical distribution of the observed X_i 's.

2. Generate an i.i.d. bootstrap sample $(X_{in}^*, \eta_{in}^*)_{1 \leq i \leq n}$ of size n , from the measure $P_n := \mathbb{P}_{n, X} \times \mathbb{P}_{n, e^\circ}$.

3. Define

$$Y_{in}^* := g(X_{in}^*)^\top \hat{\beta}_n + \eta_{in}^*,$$

for $i = 1, \dots, n$, where $\hat{\beta}_n$ is the LSE obtained in (7). Compute the bootstrapped LSE $\hat{\beta}_n^*$, using the bootstrap sample $(Y_{in}^*, X_{in}^*), i = 1, \dots, n$. Also compute the bootstrap residuals

$$e_{in}^* = Y_{in}^* - g(X_{in}^*)^\top \hat{\beta}_n^*,$$

for $i = 1, \dots, n$.

4. Compute the bootstrap test statistic T_n^* , defined as in (9), with X_i replaced by X_{in}^* , and e_i replaced by e_{in}^* , for $i = 1, \dots, n$. We approximate the distribution of nT_n by the conditional distribution of nT_n^* , given the data.

Assume that we have an infinite array of random vectors Z_1, Z_2, \dots where $Z_i := (X_i, \eta_i)$ are i.i.d. from P defined on some probability space $(\Omega, \mathcal{A}, \mathbb{P})$. We denote by \mathfrak{Z} the entire sequence $(Z_i)_{i \geq 1}$ and write $\mathbb{P}_\omega = \mathbb{P}(\cdot | \mathfrak{Z})$ and $\mathbb{E}_\omega = \mathbb{E}(\cdot | \mathfrak{Z})$ to denote conditional probability and conditional expectation, respectively, given \mathfrak{Z} .

The following result shows that under H_0 , the distribution of nT_n^* , given the data $(X_i, Y_i)_{1 \leq i \leq n}$, almost surely, converges to the same limiting distribution as that of nT_n . Thus the bootstrap is strongly consistent and we can approximate the distribution function of nT_n by $\mathbb{P}_\omega(nT_n^* \leq \cdot)$, and use it to find the one-sided cut-off for testing H_0 . To prove the result, we will need similar but slightly stronger conditions than those stated in (M). Recall that $\epsilon = m(X) - g(X)^\top \tilde{\beta}_0 + \eta$, and set $\epsilon^o := \epsilon - \mathbb{E}[\epsilon]$.

(M'') There exists $\delta > 0$ such that

$$(a) \quad \mathbb{E}[|g(X)|_\infty^{4+2\delta}] < \infty \text{ and } \mathbb{E}[|m(X)|^{2+\delta}] < \infty,$$

$$(b) \quad \mathbb{E}[|\eta|^{2+\delta}] < \infty,$$

$$(c) \quad \text{for any } 1 \leq q, r, s, t \leq 4,$$

$$\mathbb{E} \left[|k(X_q, X_r)|^{2+\delta} (1 + |g(X_s)|_\infty^{2+\delta}) (1 + |g(X_t)|_\infty^{2+\delta}) \right] < \infty,$$

$$(d) \quad \text{for any } 1 \leq q, r \leq 2,$$

$$\mathbb{E}[|l(\epsilon_q^o, \epsilon_r^o)|^{2+\delta}] < \infty,$$

$$\mathbb{E}[(1 + |g(X_q)|_\infty^{2+\delta}) |f(\epsilon_q^o, \epsilon_r^o)|^{2+\delta}] < \infty \quad \text{for } f = l_x, l_y,$$

and

$$\mathbb{E}[(1 + |g(X_q)|_\infty^{2+\delta} + |g(X_q)|_\infty^{4+2\delta}) |f(\epsilon_q^o, \epsilon_r^o)|^{2+\delta}] < \infty \quad \text{for } f = l_{xx}, l_{yy}, l_{xy}.$$

Theorem 4.1 *Suppose that conditions (I), (K) and (M'') hold. Define the random variable $\epsilon := m(X) - g(X)^\top \tilde{\beta}_0 + \eta$, and $\epsilon^o = \epsilon - \mathbb{E}[\epsilon]$, where $(X, \eta) \sim P$. Then*

$$nT_n^* \rightarrow_d \chi(P_X \times P_{\epsilon^o}),$$

conditional on the observed data a.s., where χ is given in Theorem 3.1. As a consequence, under H_0 , and $nT_n^ \rightarrow_d \chi(P_X \times P_\eta)$, conditional on the observed data a.s.*

Remark 4.1 *Note that it follows from Theorem 3.2 that $nT_n \rightarrow \infty$ in probability under H_1, H_2 or H_3 . But by Theorem 4.1, the quantiles of the conditional distribution of nT_n^* are tight. Hence, the power of our test under H_1, H_2 or H_3 converges to 1 as $n \rightarrow \infty$.*

Remark 4.2 (Gaussian kernels) *One of the natural choices for k and l is the Gaussian kernels. In this case, we can take $k(u, u') = \exp(-\sigma^{-2}\|u - u'\|^2)$ and $l(v, v') = \exp(-\gamma^{-2}|v - v'|^2)$ where $u, u' \in \mathbb{R}^{d_0}$, $v, v' \in \mathbb{R}$ and σ and γ are fixed parameters (can be taken to be 1). Then k and l satisfy condition (K). Since the Gaussian kernels are bounded with all their partial derivatives being bounded, conditions (M.d), (M'.c), (M''.c-d) are automatically satisfied for any joint distribution of (X, η) . Also, condition (M.c) is implied by the simpler condition $\mathbb{E}[|g(X)|_\infty^4] < \infty$.*

5 Simulation study

In this section we investigate the finite sample performance of the proposed testing procedure based on T_n , as defined in (9), in two different scenarios: (a) testing for the independence of the error η and the predictor X , as in (1), when the regression model is well-specified; (b) testing for the goodness-of-fit of the parametric regression model when the independence of η and X is assumed. As we have discussed in the Introduction, there are very few methods available to test (a), and hardly any when $d_0 > 2$. For the goodness-of-fit of the parametric regression model there has been quite a lot of work in the statistical literature but we only consider a selected few procedures for comparison.

We consider two data generating models. Model 1 is adapted from [SMQ98] (see Model 3 of their paper) and can be expressed as

$$Y = 2 + 5X_1 - X_2 + aX_1X_2 + \eta,$$

with covariate $X = (X_1, \dots, X_{d_0})^\top$, where X_1, \dots, X_{d_0} are distributed independently as $\text{Uniform}(0, 1)$, and η is drawn from a normal distribution with mean 0. [SMQ98] used $d_0 = 2$ in their simulations but we use $d_0 = 4$.

λ	0	5	10	15	20	25	50
$n = 100$	5	15	26	31	32	36	44
$n = 200$	5	40	68	74	78	81	88

Table 1: Percentage of times Model 1 was rejected, $\alpha = 0.05$.

λ	0	5	10	15	20	25	50
$n = 100$	6	27	30	34	34	34	37
$n = 200$	5	49	63	69	71	71	75

Table 2: Percentage of times Model 2 was rejected, $\alpha = 0.05$.

The other model, Model 2, is adapted from [FH01] (see Example 4 of their paper) and can be written as

$$Y = X_1 + aX_2^2 + 2X_4 + \eta,$$

where $X = (X_1, X_2, X_3, X_4)^\top$ is the covariate vector. The covariates X_1, X_2, X_3 are normally distributed with mean 0 and variance 1 and pairwise correlation 0.5. The predictor X_4 is binary with probability of “success” 0.4 and independent of X_1, X_2 and X_3 . Random samples of size n , where n is considered to be 100 and 200, are drawn from Model 1 (and also from Model 2) and a multiple linear regression model is fitted to the samples, without the interaction X_1X_2 term (X_2^2 term). Thus, these models are well-specified if and only if $a = 0$. In all the calculations, whenever required, we use 1000 bootstrap samples to estimate the cut-off values for the tests. To implement our procedure we take Gaussian kernels with fixed bandwidths. To make our procedure invariant under linear transformations we work with standardized variables.

5.1 Testing for the independence

We consider the above two models with $a = 0$ and the following error structure:

$$\eta|X_1 \sim N\left(0, \frac{10 + \lambda|X_1|}{10}\right),$$

where $\lambda = 0, 5, 10, 15, 20, 25, 50$. Table 1 gives the percentage of times Model 1 was rejected as the sample size n and λ vary, when $\alpha = 0.05$. Table 2 gives the same results for Model 2. As expected, the power of the test increases monotonically with an increase in λ and n .

a	0	1	2	3	4	5	7	10
T_n	4	6	11	20	34	57	89	100
S_1	5	5	7	8	11	16	28	48
S_2	3	5	7	10	14	21	37	60
F	7	7	7	10	16	22	46	89

Table 3: Percentage of times Model 1 was rejected when $\alpha = 0.05$ and $n = 100$.

a	0	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.50	0.60
T_n	6	7	10	14	22	31	43	57	69	81	92
S_1	5	5	6	8	13	15	25	31	41	51	69
S_2	5	5	8	10	16	20	31	38	49	55	68
F	8	10	8	9	10	16	23	31	42	64	84

Table 4: Percentage of times Model 2 was rejected when $\alpha = 0.05$ and $n = 100$.

5.2 Goodness-of-fit test for parametric regression

Under the assumption of independence of X and η , our procedure can be used to test the goodness-of-fit of the fitted parametric model. In our simulation study we compare the performance of our method with that of [SMQ98] and [FH01]. [FH01] proposed a lack-of-fit test based on Fourier transforms under the assumption of i.i.d. Gaussian errors; also see [CS10] for a very similar method. The main drawback of this approach is that the method needs a reliable estimator of σ^2 (the variance of η) to compute the test-statistic, and it can be very difficult to obtain such an estimator under model mis-specification. We present the power study of the adaptive Neyman test ($T_{AN,1}^*$; see equation (2.1) of [FH01]) using the *known* σ^2 (as a gold standard). We denote this test-statistic by F . Note that when using an estimate of σ^2 , as in equation (2.10) of [FH01], we got very poor results. [SMQ98] uses the empirical process of the regressors marked by the residuals and use bootstrap approximations to find the cut-off of the test statistic. We implement this method using the IntRegGOF library in the R package. We denote the two variant test statistics – the Kolmogorov-Smirnov type and the Cramér-von Mises type – by S_1 and S_2 , respectively. From Tables 3 and 4 it is clear that our procedure has much better finite sample performance when compared to the competing methods. Note that as a increasing, the power of the test monotonically increases to 1. It even performs better than F , which uses the known σ^2 .

5.3 Real data analysis

We consider two well-known regression data sets and illustrate the performance of our procedure on them. The first data set involves understanding the relation between the atmospheric ozone level and a variety of atmospheric pollutants (e.g., nitrogen dioxide, carbon dioxide, sulphur dioxide, etc.) and contain 8 predictors, and is studied in equation (2) of [Xia09]. For a complete background on the data set see the reports of the World Health Organization (2003), Bonn, Switzerland. As illustrated in [Xia09], the data set certainly exhibits a non-linear trend. However, neither [SMQ98] nor [FH01] reject the linear model specification at 5% significance level, which implies that their methods are not efficient with multiple regressors. Our procedure yields a p -value of 0.02.

The other example we consider is the *savings* data set given in [Far05] (see Chapter 3, page 31). This data set consists of some economic measurements collected for 50 countries and has 4 predictors. It is used in the book as an illustration of the various inferential techniques using multiple linear regression. For this data our method, along with that of [SMQ98] and [FH01], accepts the null hypothesis (3). We observe a p -value of 0.2 for our method. Thus, there is a natural agreement among the competing procedures, as can be expected from a data set used to demonstrate multiple linear regression.

6 A convergence result

In this section we present a result on triangular arrays of random variables that will help us understand the limit behavior of T_n under the null and alternative hypotheses and yield the consistency of our bootstrap procedure.

We denote by $Z = (X, \epsilon) \sim P$ on $\mathbb{R}^{d_0} \times \mathbb{R}$. For each $n \geq 1$, we will consider a triangular array of random vectors $Z_{in} := (X_{in}, \epsilon_{in})$ for $i = 1, \dots, n$, i.i.d. from a distribution P_n on $\mathbb{R}^{d_0} \times \mathbb{R}$ and a real vector β_n , and define

$$Y_{in} = g(X_{in})^\top \beta_n + \epsilon_{in},$$

for $i = 1, 2, \dots, n$. We may assume that the random vectors $Z, (Z_{in})_{1 \leq i \leq n < \infty}$ are all defined on a common probability space and we denote by \mathbb{P} and \mathbb{E} the probability measure and the corresponding expectation operator on that probability space.

We compute an estimator of β_n , to be denoted by β_n^* , from the given data using the method of least squares, i.e.,

$$\beta_n^* = \operatorname{argmin}_{\beta \in \mathbb{R}^d} \sum_{i=1}^n (Y_{in} - g(X_{in})^\top \beta)^2 = A_n^{-1} \cdot \frac{1}{n} \sum_{i=1}^n g(X_{in}) Y_{in},$$

where $A_n := \frac{1}{n} \sum_{i=1}^n g(X_{in})g(X_{in})^\top$ is assumed to be invertible. We write

$$\epsilon_{in}^* = Y_{in} - g(X_{in})^\top \beta_n^*$$

for the i -th residual (at stage n). We want to find the limit distribution of the test statistic

$$T_n^* = \frac{1}{n^2} \sum_{i,j} k_{ij} l_{ij}^* + \frac{1}{n^4} \sum_{i,j,q,r} k_{ij} l_{qr}^* - 2 \frac{1}{n^3} \sum_{i,j,q} k_{ij} l_{iq}^*,$$

where $k : \mathbb{R}^{d_0} \times \mathbb{R}^{d_0} \rightarrow \mathbb{R}$, $l : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ are kernels, $k_{ij} = k(X_{in}, X_{jn})$, and $l_{ij}^* = l(\epsilon_{in}^*, \epsilon_{jn}^*)$. We make the following assumptions to study the limiting behavior of T_n^* :

(C1) **On measures P_n .** We will assume the following conditions.

- (a) X_{in} and ϵ_{in} are independent. In other words, $P_n = P_{n,X} \times P_{n,\epsilon}$, for all n , where $P_{n,X}$ is a measure on \mathbb{R}^{d_0} and $P_{n,\epsilon}$ is a measure on \mathbb{R} .
- (b) $\mathbb{E}[\epsilon_{1n}] = 0$ for all n .
- (c) There exists a distribution $P = P_X \times P_\epsilon$ on $\mathbb{R}^{d_0} \times \mathbb{R}$ such that $P_n \rightarrow_d P$.
- (d) $(X_{1n}, g(X_{1n})) \rightarrow_d (X, g(X))$, where $X \sim P_X$.

(C2) **Uniform integrability conditions.** The following families of random variables are uniformly integrable for any $1 \leq p, q, r, s \leq 4$,

- (a) $\{|g(X_{pn})|_\infty^2 : n \geq 1\}$,
- (b) $\{|\epsilon_{pn}|^2 : n \geq 1\}$,
- (c) $\{k^2(X_{pn}, X_{qn})(1 + |g(X_{rn})|_\infty^2)(1 + |g(X_{sn})|_\infty^2) : n \geq 1\}$, and
- (d) $\{f^2(\epsilon_{pn}, \epsilon_{qn}) : n \geq 1\}$ for $f = l, l_x, l_y, l_{xx}, l_{yy}, l_{xy}$.

Theorem 6.1 *Suppose that conditions (I), (K), (C1) and (C2) hold. Then $nT_n^* \rightarrow_d \chi$, where χ has a non-degenerate distribution which depends upon $P = P_X \times P_\epsilon$ and is denoted by $\chi = \chi(P_X \times P_\epsilon)$ (as in Theorem 3.1).*

The proof of Theorem 6.1 is relegated to the Appendix.

7 Proofs of the main results

7.1 Proof of the Theorem 3.1

The proof is an easy consequence of Theorem 6.1, by taking $P_n \equiv P$ for all n . Under H_0 , P is in the product form $P_X \times P_\eta$ which implies (C1.a). The conditions (C1.b-d)

are also trivially satisfied. Moreover, condition (C2) is immediate from assumption (M). \square

7.2 Proof of the Theorem 4.1

We will apply Theorem 6.1 to derive the desired result by checking that assumptions (C1) and (C2) hold for each $\omega \in \Omega$, outside a set of measure zero. Note that we will apply Theorem 6.1 conditional on \mathfrak{Z} , and thus the probability and expectation operators in Theorem 6.1 are now \mathbb{P}_ω and \mathbb{E}_ω , respectively. We will apply the theorem with $\epsilon_{in} = \eta_{in}^*$, $X_{in} = X_{in}^*$, for all $1 \leq i \leq n$, and with (random) measures $P_n = P_{n,X} \times P_{n,e^o}$ where,

$$P_{n,X} = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}, \quad \text{and} \quad P_{n,e^o} = \frac{1}{n} \sum_{i=1}^n \delta_{e_i^o}.$$

Let

$$\epsilon_i := m(X_i) - g(X_i)^\top \tilde{\beta}_0 + \eta_i, \quad (11)$$

for $i = 1, \dots, n$. Then $\epsilon_1, \epsilon_2, \dots, \epsilon_n$ are i.i.d. and let P_{e^o} be the distribution of $\epsilon_i^o := \epsilon_i - \mathbb{E}[\epsilon_i]$.

Let us start by verifying (C1). By definition, $P_n = P_{n,X} \times P_{n,e^o}$ is a product measure. We take $P = P_X \times P_{e^o}$, where P_X (resp. P_{e^o}) is the common of distribution of X_i (resp. ϵ_i^o). By Lemma A.2(ii), $P_{n,e^o} \rightarrow_d P_{e^o}$ almost surely. An application of the Glivenko-Cantelli theorem yields $P_{n,X} \rightarrow_d P_X$ almost surely. Similarly, we have $(X_{1n}^*, g(X_{1n}^*)) \rightarrow_d (X, g(X))$ a.s. Also, $\mathbb{E}[\epsilon_{1n}] = \mathbb{E}_\omega[\eta_{1n}^*] = \mathbb{P}_n[e - \bar{e}] = 0$.

We will now show that (C2) holds. First note that $\mathbb{E}_\omega[|\eta_{1n}^*|^{2+\delta}] = \mathbb{P}_n[|e^o|^{2+\delta}]$ is bounded by a constant (depending on ω) by Lemma A.2(iii). This shows (C2.b). To see that (C2.a) holds, observe that by assumption (M'').a) and the SLLN, almost surely

$$\mathbb{E}_\omega[|g(X_{1n}^*)|_\infty^{2+\delta}] = \mathbb{P}_n[|g(X)|_\infty^{2+\delta}] \rightarrow \mathbb{E}[|g(X)|_\infty^{2+\delta}] < \infty.$$

To verify (C2.c), notice that the quantity of interest is a V-statistic. The SLLN for U-statistics along with the condition (M'').c) imply that (C2.c) holds.

It remains to check (C2.d). Throughout the rest of proof, we will use the notation ' $a_n \lesssim b_n$ ' for two positive sequences of real numbers a_n and b_n to mean that $a_n \leq Cb_n$, for all n for some constant C . Consider $f = l_{xx}, l_{xy}$ or l_{yy} . Then, for $q \neq r$,

$$\mathbb{E}_\omega[|f(\eta_{qn}^*, \eta_{rn}^*)|^{2+\delta}] = \frac{1}{n^2} \sum_{i,j=1}^n |f(e_i^o, e_j^o)|^{2+\delta},$$

that can be bounded by

$$\begin{aligned} & \frac{2^{2+\delta}}{n^2} \sum_{i,j=1}^n \left[|f(e_i^o, e_j^o) - f(\epsilon_i^o, \epsilon_j^o)|^{2+\delta} + |f(\epsilon_i^o, \epsilon_j^o)|^{2+\delta} \right] \\ & \lesssim \mathbb{P}_n[|e^o - \epsilon^o|^{2+\delta}] + \frac{1}{n^2} \sum_{i,j=1}^n |f(\epsilon_i^o, \epsilon_j^o)|^{2+\delta} = O_\omega(1). \end{aligned} \quad (12)$$

In the first inequality above, we have used the Lipschitz continuity of f . Note that by the SLLN of V-statistics $n^{-2} \sum_{i,j=1}^n |f(\epsilon_i^o, \epsilon_j^o)|^{2+\delta} \xrightarrow{a.s.} \mathbb{E}[|f(\epsilon_1^o, \epsilon_2^o)|^{2+\delta}]$, which holds under the moment condition $\mathbb{E}[|f(\epsilon_q^o, \epsilon_r^o)|^{2+\delta}] < \infty$ for $f = l_{xx}, l_{xy}$ or l_{yy} from (M''.d). This fact along with Lemma A.2(i) justifies the equality in (12).

A similar analysis can be done for $q = r$. Indeed, $\mathbb{E}_\omega[|f(\eta_{qn}^*, \eta_{qn}^*)|^{2+\delta}] = \sum_{i=1}^n |f(e_i^o, e_i^o)|^{2+\delta}/n$ is bounded by

$$\begin{aligned} & \frac{2^{2+\delta}}{n} \sum_{i=1}^n \left[|f(e_i^o, e_i^o) - f(\epsilon_i^o, \epsilon_i^o)|^{2+\delta} + |f(\epsilon_i^o, \epsilon_i^o)|^{2+\delta} \right] \\ & \lesssim \mathbb{P}_n[|e_i^o - \epsilon_i^o|^{2+\delta}] + \frac{1}{n} \sum_{i=1}^n |f(\epsilon_i^o, \epsilon_i^o)|^{2+\delta} = O_\omega(1). \end{aligned}$$

Now consider $f = l_x$ or l_y . For $1 \leq i \leq n$, let $a_i := |e_i^o - \epsilon_i^o|$. Consider the following upper bound for $|f(e_i^o, e_j^o)|$ which uses one term Taylor expansion for f and the Lipschitz continuity of the partial derivatives f_x and f_y .

$$|f(e_i^o, e_j^o)| \leq |f(\epsilon_i^o, \epsilon_j^o)| + a_i |f_x(\epsilon_i^o, \epsilon_j^o)| + a_j |f_y(\epsilon_i^o, \epsilon_j^o)| + 2L(a_i + a_j), \quad (13)$$

Consequently, if $q \neq r$, $\mathbb{E}_\omega[|f(\eta_{qn}^*, \eta_{rn}^*)|^{2+\delta}]$ is bounded by the following, up to a constant,

$$\frac{1}{n^2} \sum_{i,j} |f(\epsilon_i^o, \epsilon_j^o)|^{2+\delta} + \frac{1}{n^2} \sum_{i,j} \left[|a_i f_x(\epsilon_i^o, \epsilon_j^o)|^{2+\delta} + |a_j f_y(\epsilon_i^o, \epsilon_j^o)|^{2+\delta} \right] + \mathbb{P}_n[|a|^{2+\delta}].$$

The first and the third term are $O_\omega(1)$ by (M''.d) and Lemma A.2(i). Note that

$$a_i \leq d|\hat{\beta}_n - \tilde{\beta}_0|_\infty |g(X_i)|_\infty + |\bar{e} - \mathbb{E}[\epsilon]| = O_\omega(1)(1 + |g(X_i)|_\infty).$$

Therefore,

$$\frac{1}{n^2} \sum_{i,j} |a_i f_x(\epsilon_i^o, \epsilon_j^o)|^{2+\delta} \leq O_\omega(1) \cdot \frac{1}{n^2} \sum_{i,j} |(1 + |g(X_i)|_\infty) f_x(\epsilon_i^o, \epsilon_j^o)|^{2+\delta},$$

which is again $O_\omega(1)$ by the SLLN of V-statistics which holds under (M''.d). Similarly,

$$\frac{1}{n^2} \sum_{i,j} |a_j f_y(\epsilon_i^o, \epsilon_j^o)|^{2+\delta} = O_\omega(1).$$

Putting these together, we obtain that

$$\mathbb{E}_\omega[|f(\eta_{qn}^*, \eta_{rn}^*)|^{2+\delta}] = O_\omega(1) \quad \text{for } q \neq r.$$

A similar analysis shows that $\mathbb{E}_\omega[|f(\eta_{qn}^*, \eta_{qn}^*)|^{2+\delta}] = O_\omega(1)$.

For $f = l$, we can closely imitate the above argument for $f = l_x$ or l_y to deduce that $\mathbb{E}_\omega[|f(\eta_{qn}^*, \eta_{rn}^*)|^{2+\delta}] = O_\omega(1)$ for any $1 \leq q, r \leq 2$. We just need to replace the inequality (13) with the following inequality that now follows from the two-term Taylor expansion of the function l .

$$\begin{aligned} |l(e_i^o, e_j^o)| &\leq |l(\epsilon_i^o, \epsilon_j^o)| + a_i |l_x(\epsilon_i^o, \epsilon_j^o)| + a_j |l_y(\epsilon_i^o, \epsilon_j^o)| + \frac{1}{2} a_i^2 |l_{xx}(\epsilon_i^o, \epsilon_j^o)| \\ &\quad + \frac{1}{2} a_j^2 |l_{yy}(\epsilon_i^o, \epsilon_j^o)| + a_i a_j |l_{xy}(\epsilon_i^o, \epsilon_j^o)| + 4L(a_i^2 + a_j^2), \end{aligned}$$

We omit the details. Thus assumption (C2.d) of Theorem 6.1 holds. This concludes the proof of the Theorem 4.1. \square

7.3 Proof of Theorem 3.2

Let ϵ_i be as defined in (11). Note that the LSE $\hat{\beta}_n$ as defined in (7) admits the following expansion around $\tilde{\beta}_0$:

$$\begin{aligned} \sqrt{n}(\hat{\beta}_n - \tilde{\beta}_0) &= \sqrt{n} \left(A_n^{-1} \frac{1}{n} \sum_{i=1}^n g(X_i) Y_i - \tilde{\beta}_0 \right) \\ &= \sqrt{n} \left(A_n^{-1} \frac{1}{n} \sum_{i=1}^n g(X_i) (m(X_i) - g(X_i)^\top \tilde{\beta}_0 + \eta_i) \right), \\ &= (I + o_{\mathbb{P}}(1)) n^{-1/2} \sum_{i=1}^n A^{-1} g(X_i) \epsilon_i, \end{aligned} \tag{14}$$

where in the last step we have used the fact that $A_n \xrightarrow{a.s.} A$, which holds as $\mathbb{E}[|g(X)|_\infty^2] < \infty$. By (14) $\hat{\beta}_n$ admits the following expansion around $\tilde{\beta}_0$:

$$\sqrt{n}(\hat{\beta}_n - \tilde{\beta}_0) = (I + o_{\mathbb{P}}(1)) n^{-1/2} \sum_{i=1}^n A^{-1} g(X_i) \epsilon_i.$$

The normal equation for the regression model yields

$$\mathbb{E}[g(X)(m(X) - g(X)^\top \tilde{\beta}_0)] = 0.$$

Also, $\mathbb{E}[g(X)\eta] = \mathbb{E}[g(X)\mathbb{E}[\eta|X]] = 0$. Hence, we have $\mathbb{E}[g(X)\epsilon] = 0$. Moreover, by (M'.b), the covariance matrix $A^{-1}\mathbb{E}[g(X)g(X)^\top \epsilon^2]A^{-1}$ exists. So, by the central limit theorem (CLT), $\sqrt{n}(\hat{\beta}_n - \tilde{\beta}_0)$ converges in distribution to a Gaussian random vector with mean 0 and covariance $A^{-1}\mathbb{E}[g(X)g(X)^\top \epsilon^2]A^{-1}$.

We expand $l_{ij} := l(e_i, e_j)$ around $l(\epsilon_i, \epsilon_j)$ using Taylor's theorem as

$$l_{ij} = l(\epsilon_i, \epsilon_j) + \left[(e_i - \epsilon_i)l_x(\gamma_{ijn}, \tau_{ijn}) + (e_j - \epsilon_j)l_y(\gamma_{ijn}, \tau_{ijn}) \right]$$

for some point $(\gamma_{ijn}, \tau_{ijn})$ on the line joining (e_i, e_j) and (ϵ_i, ϵ_j) . Using

$$e_i - \epsilon_i = -g(X_i)^\top (\hat{\beta}_n - \tilde{\beta}_0), \quad (15)$$

we can decompose T_n in the following way,

$$T_n = T_n^{(0)} + (\hat{\beta}_n - \tilde{\beta}_0)^\top T_n^{(1)} + R_n,$$

where

$$T_n^{(p)} = \frac{1}{n^2} \sum_{i,j}^n k_{ij} l_{ij}^{(p)} + \frac{1}{n^4} \sum_{i,j,q,r}^n k_{ij} l_{qr}^{(p)} - 2 \frac{1}{n^3} \sum_{i,j,q}^n k_{ij} l_{iq}^{(p)},$$

for $p = 0, 1$ and

$$l_{ij}^{(0)} = l(\epsilon_i, \epsilon_j), \quad l_{ij}^{(1)} = -\left[l_x(\epsilon_i, \epsilon_j)g(X_i) + l_y(\epsilon_i, \epsilon_j)g(X_j) \right].$$

We will first show the negligibility of the reminder term R_n . More precisely, we claim that $\sqrt{n}R_n \xrightarrow{\mathbb{P}} 0$. To prove the claim we need the following elementary lemma which we state without proof.

Lemma 7.1 *Let $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ be a continuously differentiable function with its partial derivatives f_x, f_y being Lipschitz continuous with Lipschitz constant L (w.r.t. ℓ_∞ norm). Then for any $u, v \in \mathbb{R}^2$,*

$$|f(v) - f(u)| \leq 2|\nabla f(u)|_\infty |u - v|_\infty + 2L|u - v|_\infty^2.$$

An application of the above lemma together with (15) gives

$$\begin{aligned} |l_x(\gamma_{ijn}, \tau_{ijn}) - l_x(\epsilon_i, \epsilon_j)|_\infty &\lesssim |\nabla l_x(\epsilon_i, \epsilon_j)|_\infty |\hat{\beta}_n - \tilde{\beta}_0|_\infty (|g(X_i)|_\infty + |g(X_j)|_\infty) \\ &\quad + |\hat{\beta}_n - \tilde{\beta}_0|_\infty^2 (|g(X_i)|_\infty + |g(X_j)|_\infty)^2. \end{aligned}$$

Similarly, we can bound $|l_y(\gamma_{ijn}, \tau_{ijn}) - l_y(\epsilon_i, \epsilon_j)|_\infty$. Finally, we can bound $\sqrt{n}|R_n|$, up to a constant, by

$$\sqrt{n}|\hat{\beta}_n - \tilde{\beta}_0|_\infty^2 T_n^{(2)} + \sqrt{n}|\hat{\beta}_n - \tilde{\beta}_0|_\infty^3 T_n^{(3)}, \quad (16)$$

where, $T_n^{(2)}$ and $T_n^{(3)}$ are defined as follows:

$$T_n^{(p)} := \frac{1}{n^4} \sum_{i,j,q,r}^n |k_{ij}| (l_{ij}^{(p)} + l_{qr}^{(p)} + l_{iq}^{(p)}), \quad \text{for } p = 2, 3,$$

with

$$\begin{aligned} l_{ij}^{(2)} &= |\text{Hess}(l)(\epsilon_i, \epsilon_j)|_\infty (|g(X_i)|_\infty^2 + |g(X_j)|_\infty^2), \\ l_{ij}^{(3)} &= |g(X_i)|_\infty^3 + |g(X_j)|_\infty^3. \end{aligned}$$

Note that $T_n^{(2)}$ and $T_n^{(3)}$ are V-statistics whose kernels are integrable by (M'.c-iii) and (M'.c-iv). Consequently, the WLLN of V-statistics holds for $T_n^{(2)}$ and $T_n^{(3)}$. Now since $\sqrt{n}|\hat{\beta}_n - \tilde{\beta}_0|_\infty = O_{\mathbb{P}}(1)$, it follows that (16) $\xrightarrow{\mathbb{P}} 0$ and the claim is established.

Thus it remains to find the limiting distribution of $T_n^{(0)} + (\hat{\beta}_n - \tilde{\beta}_0)^\top T_n^{(1)}$. To do that first we will show that X and ϵ are not independent under each of H_1, H_2 and H_3 and hence $\theta(X, \epsilon) > 0$ where $\theta(X, \epsilon)$ is as defined in (5). Under hypothesis H_1 , $X \not\perp \eta$ and $\epsilon = \eta$. Hence $X \not\perp \epsilon$ under H_1 . For the case H_2 and H_3 we proceed as follows. The conditional mean of ϵ given X is

$$\mathbb{E}[\epsilon|X] = m(X) - g(X)^\top \tilde{\beta}_0 + \mathbb{E}[\eta|X] = m(X) - g(X)^\top \tilde{\beta}_0.$$

Note that under H_2 or H_3 , $m(X) \neq g(X)^\top \tilde{\beta}_0$ with positive probability. In the case when $m(X) - g(X)^\top \tilde{\beta}_0$ is a non-constant function of X , we have that $\mathbb{E}[\epsilon|X]$ depends on X , and hence X and ϵ are not independent. The case $m(X) = g(X)^\top \tilde{\beta}_0 + c$ for some non-zero constant c does not arise for H_2 by the assumption in Theorem 3.2. On the other hand, under H_3 , if $m(X) = g(X)^\top \tilde{\beta}_0 + c$, then $\epsilon = c + \eta$. Thus ϵ and X are not independent.

Letting $W_i := (X_i, \epsilon_i)$, $T_n^{(p)}$ for $p = 0, 1$, can naturally be written as a V-statistic

$$T_n^{(p)} = \frac{1}{n^4} \sum_{1 \leq q, r, s, t \leq n} h^{(p)}(W_q, W_r, W_s, W_t),$$

for some symmetric kernel $h^{(p)}$ given by

$$h^{(p)}(W_q, W_r, W_s, W_t) = \frac{1}{4!} \sum_{(i,j,u,v)}^{(q,r,s,t)} k_{ij} l_{ij}^{(p)} + k_{ij} l_{uv}^{(p)} - 2k_{ij} l_{iu}^{(p)},$$

where the sum is being taken over all $4!$ permutations of (q, r, s, t) . Note that $\mathbb{E}[|h^{(0)}(W_q, W_r, W_s, W_t)|^2] < \infty$ for $1 \leq q, r, s, t \leq 4$ by (M'.c-i) under hypothesis H_j for each $j = 1, 2, 3$. Also, $\mathbb{E}[h^{(0)}(W_1, W_2, W_3, W_4)] = \theta(X, \epsilon)$ by the definition of θ . Thus appealing to the standard theory of V-statistics, we obtain

$$\sqrt{n}(T_n^{(0)} - \theta(X, \epsilon)) = n^{-1/2} \sum_{i=1}^n h_1^{(0)}(W_i) + o_{\mathbb{P}}(1), \quad (17)$$

where $h_1^{(0)}(w) = \mathbb{E}[h^{(0)}(w, W_2, W_3, W_4)] - \theta(X, \epsilon)$. Note that $\mathbb{E}[h_1^{(0)}(W_i)] = 0$ and $\mathbb{E}[h_1^{(0)}(W_i)^2] \leq \text{Var}(h^{(0)}(W_1, W_2, W_3, W_4)) < \infty$.

On the other hand, $\mathbb{E}[|h^{(1)}(W_q, W_r, W_s, W_t)|_\infty] < \infty$ for $1 \leq q, r, s, t \leq 4$ by (M'.c-ii) under hypothesis H_j for each $j = 1, 2, 3$. So by the WLLN for V-statistics, we have

$$T_n^{(1)} \xrightarrow{\mathbb{P}} \gamma := \mathbb{E}[h^{(1)}(W_1, W_2, W_3, W_4)].$$

From (14) and (17),

$$\begin{aligned} \sqrt{n}(T_n - \theta(X, \epsilon)) &= \sqrt{n}(T_n^{(0)} - \theta(X, \epsilon)) + \sqrt{n}(\hat{\beta}_n - \tilde{\beta}_0)^\top T_n^{(1)} + o_{\mathbb{P}}(1) \\ &= n^{-1/2} \sum_{i=1}^n \left[h_1^{(0)}(W_i) + \gamma^\top A^{-1} g(X_i) \epsilon_i \right] + o_{\mathbb{P}}(1), \end{aligned}$$

which by the CLT has an asymptotic normal distribution with mean 0 and variance

$$\text{Var}(h_1^{(0)}(W_1) + \gamma^\top A^{-1} g(X_1) \epsilon_1). \quad (18)$$

□

Acknowledgments. The authors would like to thank Probal Chaudhuri, Victor de la Pena, Bharath Sriperumbudur and Gábor Székely for helpful discussions.

A Appendix

A.1 Proof of the Theorem 6.1

Observe that

$$\epsilon_{in}^* - \epsilon_{in} = -(\beta_n^* - \beta_n)^\top g(X_{in}). \quad (19)$$

Using (19) and by Taylor's expansion

$$l_{ij}^* = l_{ij}^{(0)} + (\beta_n^* - \beta_n)^\top l_{ij}^{(1)} + \frac{1}{2}(\beta_n^* - \beta_n)^\top v_{ij}^* (\beta_n^* - \beta_n) \quad (20)$$

where

$$\begin{aligned} l_{ij}^{(0)} &:= l_{ij} = l(\epsilon_{in}, \epsilon_{jn}), \quad l_{ij}^{(1)} := -\left[l_x(\epsilon_{in}, \epsilon_{jn}) g(X_{in}) + l_y(\epsilon_{in}, \epsilon_{jn}) g(X_{jn}) \right], \text{ and} \\ v_{ij}^* &:= \left[l_{xx}(\vartheta_{ijn}, \tau_{ijn}) g(X_{in}) g(X_{in})^\top + l_{yy}(\vartheta_{ijn}, \tau_{ijn}) g(X_{in}) g(X_{in})^\top \right. \\ &\quad \left. + 2l_{xy}(\vartheta_{ijn}, \tau_{ijn}) g(X_{in}) g(X_{jn})^\top \right], \end{aligned}$$

for some point $(\vartheta_{ijn}, \tau_{ijn})$ on the straight line connecting the two points $(\epsilon_{in}^*, \epsilon_{jn}^*)$ and $(\epsilon_{in}, \epsilon_{jn})$.

In view of (20), we can decompose T_n^* in the following way

$$T_n^* = T_n^{(0)} + (\beta_n^* - \beta_n)^\top T_n^{(1)} + \frac{1}{2}(\beta_n^* - \beta_n)^\top T_n^{(2)} (\beta_n^* - \beta_n) + R_n, \quad (21)$$

where

$$T_n^{(p)} = \frac{1}{n^2} \sum_{i,j}^n k_{ij} l_{ij}^{(p)} + \frac{1}{n^4} \sum_{i,j,q,r}^n k_{ij} l_{qr}^{(p)} - \frac{2}{n^3} \sum_{i,j,q}^n k_{ij} l_{iq}^{(p)},$$

for $p \in \{0, 1, 2\}$, and

$$l_{ij}^{(2)} = \left[l_{xx}(\epsilon_{in}, \epsilon_{jn}) g(X_{in}) g(X_{in})^\top + l_{yy}(\epsilon_{in}, \epsilon_{jn}) g(X_{jn}) g(X_{jn})^\top + 2l_{xy}(\epsilon_{in}, \epsilon_{jn}) g(X_{in}) g(X_{jn})^\top \right],$$

and R_n is the reminder term. Note that $l_{ij}^{(0)} \in \mathbb{R}$, $l_{ij}^{(1)} \in \mathbb{R}^d$, and $l_{ij}^{(2)} \in \mathbb{R}^{d \times d}$.

For $p \in \{0, 1, 2\}$, $T_n^{(p)}$ can be expressed as a V -statistic (but with triangular arrays) of the form

$$T_n^{(p)} = \frac{1}{n^4} \sum_{1 \leq i,j,q,r \leq n} h^{(p)}(Z_{in}, Z_{jn}, Z_{qn}, Z_{rn}), \quad (22)$$

for some symmetric kernel $h^{(p)}$ given by

$$h^{(p)}(Z_{in}, Z_{jn}, Z_{qn}, Z_{rn}) = \frac{1}{4!} \sum_{(t,u,v,w)}^{(i,j,q,r)} k_{tu} l_{tu}^{(p)} + k_{tu} l_{vw}^{(p)} - 2k_{tu} l_{tv}^{(p)}, \quad (23)$$

where the sum is being taken over all $4!$ permutations of (i, j, q, r) .

A.1.1 Getting rid of triangular sequence

Let $Z_i := (X_i, \epsilon_i)$ be i.i.d. from P . By the Skorohod representation theorem, there exists a sufficiently rich probability space $(\tilde{\Omega}, \tilde{P})$, independent random elements $\omega_1, \omega_2, \dots$ defined on $\tilde{\Omega}$ and functions f_n, f with $\tilde{Z}_{in} := f_n(\omega_i)$, $\tilde{Z}_i := f(\omega_i)$ such that $\tilde{Z}_{in} \xrightarrow{d} Z_{in}$, $\tilde{Z}_i \xrightarrow{d} Z_i$ and

$$\tilde{Z}_{in} \xrightarrow{\tilde{P}\text{-a.s.}} \tilde{Z}_i, \text{ as } n \rightarrow \infty.$$

Since we are only concerned about the distributional limit of nT_n^* , henceforth in this proof, we may assume, without loss of generality, that for each n , the random vectors $W_{in} := (Z_{in}, Z_i)$ are independent and for each i , $Z_{in} \rightarrow Z_i$ almost surely as $n \rightarrow \infty$. This argument is similar to that in [LN09].

We will start by showing that

$$A_n := n^{-1} \sum_{i=1}^n g(X_{in}) g(X_{in})^\top \xrightarrow{\mathbb{P}} A := \mathbb{E}[g(X_1) g(X_1)^\top].$$

By assumption (C2.a), for any $1 \leq p, q \leq d$, $g_p(X_{1n})g_q(X_{1n})$ are uniformly integrable. Moreover, by (C1.d), we have $g_p(X_{1n})g_q(X_{1n}) \rightarrow_d g_p(X_1)g_q(X_1)$. Therefore, $g_p(X_{1n})g_q(X_{1n}) \xrightarrow{L^1} g_p(X_1)g_q(X_1)$ and $\mathbb{E}[|g_p(X_1)g_q(X_1)|] < \infty$. Hence, we obtain that $n^{-1} \sum_{i=1}^n g(X_i)g(X_i)^\top \xrightarrow{\mathbb{P}} A$ by the WLLN. Finally, observe that

$$n^{-1} \sum_{i=1}^n g(X_{in})g(X_{in})^\top - n^{-1} \sum_{i=1}^n g(X_i)g(X_i)^\top \xrightarrow{L^1} 0,$$

as $n \rightarrow \infty$ since $g_p(X_{1n})g_q(X_{1n}) \xrightarrow{L^1} g_p(X_1)g_q(X_1)$. This completes the proof that $A_n \xrightarrow{\mathbb{P}} A$. As a consequence, A_n is invertible (and hence β_n^* is defined) with high probability as $n \rightarrow \infty$.

Note that β_n^* admits the following expansion

$$\begin{aligned} n^{1/2}(\beta_n^* - \beta_n) &= n^{-1/2} A_n^{-1} \sum_{i=1}^n g(X_{in}) (Y_{in} - g(X_{in})^\top \beta_n) \\ &= n^{-1/2} A_n^{-1} \sum_{i=1}^n g(X_{in}) \epsilon_{in}. \end{aligned} \quad (24)$$

Next we claim that

$$n^{1/2}(\beta_n^* - \beta_n) - \zeta_n \xrightarrow{\mathbb{P}} 0, \quad (25)$$

where $\zeta_n := n^{-1/2} A^{-1} \sum_{i=1}^n g(X_i) \epsilon_i$. We will first show that

$$n^{-1/2} A^{-1} \sum_{i=1}^n g(X_{in}) \epsilon_{in} - \zeta_n \xrightarrow{L^2} 0. \quad (26)$$

Clearly, it suffices to show that $n^{-1/2} \sum_{i=1}^n (g_p(X_{in}) \epsilon_{in} - g_p(X_i) \epsilon_i) \xrightarrow{L^2} 0$ for each $1 \leq p \leq d$. Indeed, we can write the square of its L^2 -norm as

$$\begin{aligned} n^{-1} \mathbb{E} \left[\sum_{i,j=1}^n (g_p(X_{in}) \epsilon_{in} - g_p(X_i) \epsilon_i) (g_p(X_{jn}) \epsilon_{jn} - g_p(X_j) \epsilon_j) \right] \\ = \mathbb{E} \left[(g_p(X_{1n}) \epsilon_{1n} - g_p(X_1) \epsilon_1)^2 \right], \end{aligned}$$

which goes to 0 as $n \rightarrow \infty$.

This is because $g_p(X_{1n}) \epsilon_{1n} \rightarrow_d g_p(X_1) \epsilon_1$ and $g_p^2(X_{1n}) \epsilon_{1n}^2$ is uniformly integrable by assumption (C2.a), (C2.b) and the independence of X_{1n} and ϵ_{1n} . This proves (26). Recall that, from (24),

$$n^{1/2}(\beta_n^* - \beta_n) = (A_n^{-1} A) \times n^{-1/2} A^{-1} \sum_{i=1}^n g(X_{in}) \epsilon_{in}.$$

Since by CLT, ζ_n converges in distribution to a multivariate normal, (26) implies that $n^{-1/2}A^{-1}\sum_{i=1}^n g(X_{in})\epsilon_{in} = O_{\mathbb{P}}(1)$. Consequently,

$$n^{1/2}(\beta_n^* - \beta_n) - n^{-1/2}A^{-1}\sum_{i=1}^n g(X_{in})\epsilon_{in} \xrightarrow{\mathbb{P}} 0.$$

Now (25) follows from (26).

Let $V_n^{(p)}$, for $p \in \{0, 1, 2\}$, be defined analogously as $T_n^{(p)}$ in (22) but with $Z_{in} = (X_{in}, \epsilon_{in})$ replaced by $Z_i = (X_i, \epsilon_i)$. Note that $V_n^{(p)}$ is a proper V-statistic. Our next goal is to show that

$$n^{1-p/2}(T_n^{(p)} - V_n^{(p)}) \xrightarrow{L^2} 0, \quad \text{for } p \in \{0, 1, 2\}. \quad (27)$$

Observe that,

$$\mathbb{E} \left[n^{2-p} \text{tr}((T_n^{(p)} - V_n^{(p)})(T_n^{(p)} - V_n^{(p)})^\top) \right] = \frac{1}{n^{6+p}} \sum_{\mathbf{i}, \mathbf{j}} \mathbb{E}[\text{tr}(\bar{h}^{(p)}(\mathbf{i})\bar{h}^{(p)}(\mathbf{j})^\top)],$$

where $\mathbf{i} = (i_1, i_2, i_3, i_4)$ and $\mathbf{j} = (j_1, j_2, j_3, j_4)$ are multi-indices in $\{1, \dots, n\}^4$, and

$$\bar{h}^{(p)}(\mathbf{i}) := h^{(p)}(Z_{i_1n}, \dots, Z_{i_4n}) - h^{(p)}(Z_{i_1}, \dots, Z_{i_4}).$$

We will first show that $|h^{(p)}(Z_{i_1n}, \dots, Z_{i_4n})|_\infty^2$ is uniformly integrable. It is enough to show that each of the terms like $|k_{rs}l_{tu}^{(p)}|_\infty^2$, where $r, s, t, u \in \{1, 2, 3, 4\}$ may not be necessarily distinct, is uniformly integrable. Using the independence of X_{in} and ϵ_{in} , we see that this follows directly from assumption (C2.c) and (C2.d). Assumption (C1.d) and the continuous mapping theorem implies that,

$$h^{(p)}(Z_{i_1n}, \dots, Z_{i_4n}) \rightarrow_d h^{(p)}(Z_{i_1}, \dots, Z_{i_4}).$$

Thus the above convergence also holds in L^2 and we have that

$$\mathbb{E}[|h^{(p)}(Z_{i_1}, \dots, Z_{i_4})|_\infty^2] < \infty.$$

Consequently, $\mathbb{E}[|\bar{h}^{(p)}(\mathbf{i})|_\infty^2]$ is uniformly bounded for all \mathbf{i} and n . An application of the Cauchy-Schwarz inequality yields

$$\mathbb{E}[|\bar{h}^{(p)}(\mathbf{i})|_\infty |\bar{h}^{(p)}(\mathbf{j})|_\infty] \leq (\mathbb{E}[|\bar{h}^{(p)}(\mathbf{i})|_\infty^2])^{1/2} (\mathbb{E}[|\bar{h}^{(p)}(\mathbf{j})|_\infty^2])^{1/2},$$

implying that $\mathbb{E}[|\bar{h}^{(p)}(\mathbf{i})|_\infty |\bar{h}^{(p)}(\mathbf{j})|_\infty]$ is uniformly bounded. We now make the following observations.

1. The number of multi-indices \mathbf{i} and \mathbf{j} for which $|\mathbf{i} \cup \mathbf{j}| = k$ is bounded above by n^k , for $1 \leq k \leq 8$.

2. The kernel $h^{(0)}$ is degenerate of order 1, hence $\mathbb{E}[\bar{h}^{(0)}(\mathbf{i})\bar{h}^{(0)}(\mathbf{j})] = 0$ when $|\mathbf{i} \cup \mathbf{j}| = 7$ or 8.
3. The kernel $\mathbb{E}[h^{(1)}(Z_{1n}, \dots, Z_{4n})] = 0$ (we will show this in Lemma A.1), hence when $|\mathbf{i} \cup \mathbf{j}| = 8$, $\mathbb{E}[\bar{h}^{(1)}(\mathbf{i})\bar{h}^{(1)}(\mathbf{j})^\top] = \mathbb{E}[\bar{h}^{(1)}(\mathbf{i})]\mathbb{E}[\bar{h}^{(1)}(\mathbf{j})^\top] = 0$.

Putting the above observations together, it remains to prove that

$$\mathbb{E}[\text{tr}(\bar{h}^{(p)}(\mathbf{i})\bar{h}^{(p)}(\mathbf{j})^\top)] \rightarrow 0 \quad \text{for any } \mathbf{i}, \mathbf{j} \text{ such that } |\mathbf{i} \cup \mathbf{j}| = 6 + p,$$

for $p \in \{0, 1, 2\}$. But this immediately follows from the fact

$$\bar{h}^{(p)}(\mathbf{i}) \xrightarrow{L^2} 0,$$

which has been already shown. Hence (27) is proved.

Finally, we claim that

$$\begin{aligned} n \times \left[T_n^{(0)} + (\beta_n^* - \beta_n)^\top T_n^{(1)} + \frac{1}{2}(\beta_n^* - \beta_n)^\top T_n^{(2)}(\beta_n^* - \beta_n) - V_n^{(0)} \right. \\ \left. - n^{-1/2} \zeta_n^\top V_n^{(1)} - \frac{1}{2} n^{-1} \zeta_n^\top V_n^{(2)} \zeta_n \right] \xrightarrow{\mathbb{P}} 0, \end{aligned} \quad (28)$$

which now easily follows from (26) and (27).

A.1.2 Negligibility of the reminder term R_n

In this subsection, we will show that the reminder term can be ignored for future analysis. More precisely, we will prove that

$$nR_n \xrightarrow{\mathbb{P}} 0. \quad (29)$$

Let us define

$$Q_n = \frac{1}{n^2} \sum_{i,j}^n k_{ij}(v_{ij}^* - l_{ij}^{(2)}) + \frac{1}{n^4} \sum_{i,j,q,r}^n k_{ij}(v_{qr}^* - l_{qr}^{(2)}) - 2 \frac{1}{n^3} \sum_{i,j,q}^n k_{ij}(v_{iq}^* - l_{iq}^{(2)}),$$

so that $R_n = \frac{1}{2}(\beta_n^* - \beta_n)^\top Q_n(\beta_n^* - \beta_n)$. Since by (25) $n^{1/2}(\beta_n^* - \beta_n) = O_{\mathbb{P}}(1)$, it is enough to show that for each $1 \leq s, t \leq d$,

$$(Q_n)_{st} \xrightarrow{\mathbb{P}} 0.$$

Note that $(Q_n)_{st}$ is the sum of three terms and each of those terms can be shown to converge to 0 in probability. We will only spell out the details for the first term leaving the other two terms for the reader.

Thus we need to show that

$$\frac{1}{n^2} \sum_{i,j}^n k_{ij} (v_{ij}^* - l_{ij}^{(2)})_{st} \xrightarrow{\mathbb{P}} 0. \quad (30)$$

Note that $(v_{ij}^* - l_{ij}^{(2)})_{st}$ can be further broken down into three terms; the first one being $(l_{xx}(\vartheta_{ijn}, \tau_{ijn}) - l_{xx}(\epsilon_{in}, \epsilon_{jn}))g_s(X_{in})g_t(X_{in})$. The other two terms involve l_{yy} and l_{xy} . Using the Lipschitz continuity of l_{xx}, l_{yy} and l_{xy} we obtain the following bound.

$$\begin{aligned} |(v_{ij}^* - l_{ij}^{(2)})_{st}| &\lesssim L|(\epsilon_{ij}^*, \epsilon_{ij}^*) - (\epsilon_{in}, \epsilon_{jn})|_\infty (|g(X_{in})|_\infty + |g(X_{jn})|_\infty)^2 \\ &\leq dL|\beta_n^* - \beta_n|_\infty (|g(X_{in})|_\infty + |g(X_{jn})|_\infty)^3. \end{aligned}$$

Therefore, $n^{-2} \sum_{i,j}^n |k_{ij}| |(v_{ij}^* - l_{ij}^{(2)})_{st}|$ is bounded above by

$$8dL|\beta_n^* - \beta_n|_\infty \cdot \frac{1}{n^2} \sum_{i,j=1}^n |k_{ij}| (|g(X_{in})|_\infty^3 + |g(X_{jn})|_\infty^3).$$

Observe that

$$\frac{1}{n^2} \sum_{i,j=1}^n \mathbb{E} [|k_{ij}| (|g(X_{in})|_\infty^3 + |g(X_{jn})|_\infty^3)] = O(1),$$

by assumption (C2.c) and hence (30) follows. We can apply similar techniques to control the other two terms in Q_n . Hence, $Q_n = o_{\mathbb{P}}(1)$.

A.1.3 Finding the limiting distribution

In this subsection, we will finally prove that nT_n^* converges to a non-degenerate distribution. Note that by (21), (28) and (29), it is enough to show that the random variable

$$nV_n^{(0)} + \zeta_n^\top (n^{1/2}V_n^{(1)}) + \frac{1}{2}\zeta_n^\top V_n^{(2)}\zeta_n$$

converges in distribution, where $V_n^{(p)}$, for $p \in \{0, 1, 2\}$, is defined near (27). The kernel $h^{(0)}$ is degenerate of order 1, i.e., $\mathbb{E}[h^{(0)}(z_1, Z_2, Z_3, Z_4)] = 0$ a.s. Let us define

$$h_2^{(0)}(z_1, z_2) := \mathbb{E}[h^{(0)}(z_1, z_2, Z_3, Z_4)]$$

and let $S_n^{(0)}$ be the V-statistic with kernel $h_2^{(0)}$, i.e.,

$$S_n^{(0)} = \frac{1}{n^2} \sum_{i,j=1}^n h_2^{(0)}(Z_i, Z_j).$$

By the standard theory of V-statistics,

$$n(V_n^{(0)} - S_n^{(0)}) \xrightarrow{\mathbb{P}} 0.$$

The symmetric function $h_2^{(0)}$ admits an eigenvalue decomposition

$$h_2^{(0)}(z_1, z_2) = \sum_{r=0}^{\infty} \lambda_r \varphi_r(z_1) \varphi_r(z_2)$$

where $(\varphi_r)_{r \geq 0}$ is an orthonormal basis of $L^2(\mathbb{R}^{d_0+1}, P)$ and λ_r is the eigenvalue corresponding to the eigenvector φ_r . Since $h_2^{(0)}$ is degenerate of order 1, $\lambda_0 = 0, \varphi_0 \equiv 1$. Therefore, $\mathbb{E}[\varphi_r(Z_1)] = 0$ for each $r \geq 1$. Also, $\sum_r \lambda_r^2 = \mathbb{E}[h_2^{(0)}(Z_1, Z_2)^2] < \infty$. We use the above decomposition of $h_2^{(0)}$ to express $S_n^{(0)}$ as follows

$$S_n^{(0)} = \sum_{r=1}^{\infty} \lambda_r \left(n^{-1/2} \sum_{i=1}^n \varphi_r(Z_i) \right)^2.$$

Let us now turn our attention to $V_n^{(1)}$. It is again a V-statistic whose kernel $h^{(1)}$ has mean zero, i.e., $\mathbb{E}[h^{(1)}(Z_1, Z_2, Z_3, Z_4)] = 0$ (see Lemma A.1). Therefore, if we define its first order projection by

$$h_1^{(1)}(z_1) := \mathbb{E}[h^{(1)}(z_1, Z_2, Z_3, Z_4)],$$

then

$$n^{1/2} V_n^{(1)} - n^{-1/2} \sum_{i=1}^n h_1^{(1)}(Z_i) \xrightarrow{\mathbb{P}} 0.$$

On the other hand, by the WLLN for V-statistics, we have

$$V_n^{(2)} \xrightarrow{\mathbb{P}} \mathbb{E}[h^{(2)}(Z_1, Z_2, Z_3, Z_4)] =: \Lambda \in \mathbb{R}^{d \times d}.$$

By multivariate CLT, the random vectors

$$\left(n^{-1/2} \sum_{i=1}^n \varphi_r(Z_i) \right)_{r \geq 1}, \quad n^{-1/2} \sum_{i=1}^n h_1^{(1)}(Z_i), \quad \zeta_n,$$

converge in distribution to jointly Gaussian random variables

$$\mathcal{Z} = (\mathcal{Z}_r)_{r \geq 1}, \mathcal{N} = (\mathcal{N}_i)_{1 \leq i \leq d}, \mathcal{W} = (\mathcal{W}_i)_{1 \leq i \leq d},$$

where \mathcal{Z}_r are i.i.d. $N(0, 1)$, $\mathcal{N} \sim N_d(0, \Xi)$ and $\mathcal{W} \sim N_d(0, \sigma^2 I)$, with $\sigma^2 := \mathbb{E}[\epsilon_1^2]$ and $\Xi := \mathbb{E}[h_1^{(1)}(Z_1) h_1^{(1)}(Z_1)^\top]$. Also, the covariance structure between the random variables $\mathcal{Z}_r, \mathcal{N}$ and \mathcal{W} are given by

$$\begin{aligned} \mathbb{E}[\mathcal{Z}_r \mathcal{N}] &= \mathbb{E}[\varphi_r(Z_1) h_1^{(1)}(Z_1)], \\ \mathbb{E}[\mathcal{Z}_r \mathcal{W}] &= A^{-1} \mathbb{E}[g(X_1) \epsilon_1 \varphi_r(Z_1)], \\ \mathbb{E}[\mathcal{W} \mathcal{N}^\top] &= A^{-1} \mathbb{E}[\epsilon_1 g(X_1) h_1^{(1)}(Z_1)^\top]. \end{aligned}$$

Therefore, by the continuous mapping theorem,

$$\begin{aligned} nT_n^* &= nV_n^{(0)} + \zeta_n^\top (n^{1/2}V_n^{(1)}) + \frac{1}{2}\zeta_n^\top V_n^{(2)}\zeta_n + o_{\mathbb{P}}(1) \\ &\rightarrow_d \sum_{r=1}^{\infty} \lambda_r \mathcal{Z}_r^2 + \sum_{i=1}^d \mathcal{W}_i \mathcal{N}_i + \frac{1}{2} \sum_{i,j=1}^d \mathcal{W}_i \Lambda_{ij} \mathcal{W}_j =: \chi, \end{aligned} \quad (31)$$

which concludes the proof of the theorem.

Lemma A.1 *Let $h^{(1)}$ be the symmetric kernel as defined in (23). Let Z_1, Z_2, Z_3 and Z_4 be i.i.d. random vectors with $Z_i = (X_i, \epsilon_i) \in \mathbb{R}^{d_0} \times \mathbb{R}$ where X_i and ϵ_i are independent. Then*

$$\mathbb{E}[h^{(1)}(Z_1, \dots, Z_4)] = 0.$$

Proof: We have

$$h^{(1)}(Z_1, Z_2, Z_3, Z_4) = \frac{1}{4!} \sum_{(t,u,v,w)}^{(1,2,3,4)} k_{tu} l_{tu}^{(1)} + k_{tu} l_{vw}^{(1)} - 2k_{tu} l_{tv}^{(1)},$$

where the sum is being taken over all $4!$ permutations of $(1, 2, 3, 4)$. We claim that $\mathbb{E}[k_{tu} l_{tu}^{(1)} + k_{tu} l_{vw}^{(1)} - 2k_{tu} l_{tv}^{(1)}] = 0$ for each such permutation from which the lemma would follow immediately. Recall that

$$l_{ij}^{(1)} = -\left[l_x(\epsilon_i, \epsilon_j)g(X_i) + l_y(\epsilon_i, \epsilon_j)g(X_j)\right] =: Q_{ij} + R_{ij} \text{ (say).}$$

Using the independence of X_i and ϵ_i , we obtain that

$$\begin{aligned} &\mathbb{E}[k_{tu} Q_{tu} + k_{tu} Q_{vw} - 2k_{tu} Q_{tv}] \\ &= -\mathbb{E}[k(X_t, X_u)g(X_t)]\mathbb{E}[l_x(\epsilon_t, \epsilon_u)] - \mathbb{E}[k(X_t, X_u)g(X_v)]\mathbb{E}[l_x(\epsilon_v, \epsilon_w)] \\ &\quad + 2\mathbb{E}[k(X_t, X_u)g(X_t)]\mathbb{E}[l_x(\epsilon_t, \epsilon_v)] \\ &= \mathbb{E}[l_x(\epsilon_1, \epsilon_2)]\left(\mathbb{E}[k(X_1, X_2)g(X_1)] - \mathbb{E}[k(X_1, X_2)g(X_3)]\right). \end{aligned}$$

Similarly,

$$\begin{aligned} &\mathbb{E}[k_{tu} R_{tu} + k_{tu} R_{vw} - 2k_{tu} R_{tv}] \\ &= -\mathbb{E}[k(X_t, X_u)g(X_u)]\mathbb{E}[l_y(\epsilon_t, \epsilon_u)] - \mathbb{E}[k(X_t, X_u)g(X_w)]\mathbb{E}[l_y(\epsilon_v, \epsilon_w)] \\ &\quad + 2\mathbb{E}[k(X_t, X_u)g(X_v)]\mathbb{E}[l_y(\epsilon_t, \epsilon_v)] \\ &= \mathbb{E}[l_y(\epsilon_1, \epsilon_2)]\left(\mathbb{E}[k(X_1, X_2)g(X_3)] - \mathbb{E}[k(X_1, X_2)g(X_2)]\right). \end{aligned}$$

Since k is symmetric, $\mathbb{E}[k(X_1, X_2)g(X_2)] = \mathbb{E}[k(X_1, X_2)g(X_1)]$ and since l is symmetric, we have $l_x(a, b) = l_y(b, a)$ which implies that $\mathbb{E}[l_x(\epsilon_1, \epsilon_2)] = \mathbb{E}[l_y(\epsilon_1, \epsilon_2)]$. Hence,

$$\mathbb{E}[k_{tu} Q_{tu} + k_{tu} Q_{vw} - 2k_{tu} Q_{tv}] + \mathbb{E}[k_{tu} R_{tu} + k_{tu} R_{vw} - 2k_{tu} R_{tv}] = 0,$$

and consequently, $\mathbb{E}[k_{tu} l_{tu}^{(1)} + k_{tu} l_{vw}^{(1)} - 2k_{tu} l_{tv}^{(1)}] = 0$. \square

A.1.4 The empirical distribution of the residuals

In the following lemma we gather a few standard results about the empirical distribution of the residuals for the linear regression model $Y = m(X) + \eta$.

Lemma A.2 *Under the conditions (I), (M''.a) and (M''.b), the following statements hold:*

- (i) $\mathbb{P}_n[|e^o - \epsilon^o|^r] \xrightarrow{a.s.} 0$ for each $0 < r \leq 4 + 2\delta$;
- (ii) $P_{n,\epsilon^o} \rightarrow_d \epsilon^o$ a.s.;
- (iii) $\sup_n \mathbb{P}_n[|e^o|^{2+\delta}] < \infty$ a.s.

Proof: Note that $e_i - \epsilon_i = -g(X_i)^\top (\hat{\beta}_n - \tilde{\beta}_0)$. Thus,

$$\mathbb{P}_n[|e - \epsilon|^r] \leq d|\hat{\beta}_n - \tilde{\beta}_0|_\infty^r \mathbb{P}_n[|g(X)|_\infty^r].$$

Hence, $\mathbb{P}_n[|e - \epsilon|^r] \xrightarrow{a.s.} 0$ using the facts that $\mathbb{E}[|g(X)|_\infty^{4+2\delta}] < \infty$ by (M''.a) and that $\hat{\beta}_n \xrightarrow{a.s.} \tilde{\beta}_0$ by (14) and $\mathbb{E}[|g(X)\epsilon|] < \infty$, the latter being guaranteed by (M''.a) and (M''.b). Next we obtain

$$\bar{e} = \mathbb{P}_n[e] = \mathbb{P}_n[m(X)] - \mathbb{P}_n[g(X)]^\top \hat{\beta}_n \xrightarrow{a.s.} \mathbb{E}[m(X)] - \mathbb{E}[g(X)]^\top \tilde{\beta}_0 = \mathbb{E}[\epsilon].$$

This completes the proof of (i).

Letting P_{n,ϵ^o} to be the empirical measure of $\epsilon_1^o, \epsilon_2^o, \dots, \epsilon_n^o$, we first observe that

$$\int e^{i\xi x} dP_{n,\epsilon^o}(x) = e^{-i\xi \bar{e}} \mathbb{P}_n[e^{i\xi e}]$$

and hence, for any $\xi \in \mathbb{R}$, we have,

$$\begin{aligned} \left| \int e^{i\xi x} dP_{n,\epsilon^o}(x) - e^{-i\xi(\bar{e} - \mathbb{E}[\epsilon])} \int e^{i\xi x} dP_{n,\epsilon^o}(x) \right| &= \left| \mathbb{P}_n[e^{i\xi e}] - \mathbb{P}_n[e^{i\xi \epsilon}] \right| \\ &\leq |\xi| \mathbb{P}_n[|e - \epsilon|] \xrightarrow{a.s.} 0, \end{aligned}$$

by applying part (i) with $r = 1$. Now since $P_{n,\epsilon^o} \rightarrow_d \epsilon^o$ a.s. by the Glivenko-Cantelli lemma and $\bar{e} \xrightarrow{a.s.} \mathbb{E}[\epsilon]$ as shown in the part (i) of the lemma, we have $\int e^{i\xi x} dP_{n,\epsilon^o}(x) \xrightarrow{a.s.} \int e^{i\xi x} dP_{\epsilon^o}(x)$, which, by Levy's continuity theorem, yields (ii).

To prove (iii), we write

$$\begin{aligned} \mathbb{P}_n[|e^o|^{2+\delta}] &= \mathbb{P}_n[|e - \bar{e}|^{2+\delta}] = \mathbb{P}_n[|(e - \epsilon) + (\epsilon - \mathbb{E}[\epsilon]) - (\bar{e} - \mathbb{E}[\epsilon])|^{2+\delta}] \\ &\leq 3^{2+\delta} \left(\mathbb{P}_n[|e - \epsilon|^{2+\delta}] + \mathbb{P}_n[|\epsilon^o|^{2+\delta}] + |\bar{e} - \mathbb{E}[\epsilon]|^{2+\delta} \right). \end{aligned}$$

The result is then an immediate consequence of the fact that $\mathbb{P}_n[|\epsilon^o|^{2+\delta}] \xrightarrow{a.s.} \mathbb{E}[|\epsilon^o|^{2+\delta}] < \infty$ by (M''.a) and (M''.b), that $\bar{e} \xrightarrow{a.s.} \mathbb{E}[\epsilon]$, and part (i) of the lemma. \square

References

- [AB93] A. Azzalini and A. Bowman, *On the use of nonparametric regression for checking linear relationships*, J. Roy. Statist. Soc. Ser. B (1993), 549–557.
- [Bie90] H.J. Bierens, *A consistent conditional moment test of functional form*, Econometrica **58** (1990), 1443–1458.
- [BP79] T.S. Breusch and A.R. Pagan, *Simple test for heteroscedasticity and random coefficient variation*, Econometrica **47** (1979), 1287–1294.
- [CKWY88] D. Cox, E. Koh, G. Wahba, and B.S. Yandell, *Testing the (parametric) null model hypothesis in (semiparametric) partial and generalized spline models*, Ann. Statist. **16** (1988), 113–119.
- [CS10] R. Christensen and S. K. Sun, *Alternative goodness-of-fit tests for linear models*, J. Amer. Statist. Assoc. **105** (2010), 291–301.
- [CW83] R. D. Cook and S. Weisberg, *Diagnostics for heteroscedasticity in regression*, Biometrika **70** (1983), 1–10.
- [ES90] R.L. Eubank and C.H. Spiegelman, *Testing the goodness of fit of a linear model via nonparametric regression techniques*, J. Amer. Statist. Assoc. **85** (1990), no. 410, 387–392.
- [EVK08a] J.H.J. Einmahl and I. Van Keilegom, *Specification tests in nonparametric regression*, J. Econometrics **143** (2008), 88–102.
- [EVK08b] ———, *Tests for independence in nonparametric regression*, Statist. Sinica **18** (2008), 601.
- [Far05] Julian J. Faraway, *Linear models with R*, Chapman & Hall/CRC Texts in Statistical Science Series, Chapman & Hall/CRC, 2005.
- [FH01] J. Fan and L.S. Huang, *Goodness-of-fit tests for parametric regression models*, J. Amer. Statist. Assoc. **96** (2001), 640–652.
- [GBSS05] A. Gretton, O. Bousquet, A.J. Smola, and B. Schölkopf, *Measuring statistical dependence with hilbert-schmidt norms*, Proceedings of the Conference on Algorithmic Learning Theory (ALT) (2005), 63–77.

- [GFT⁺08] Arthur Gretton, Kenji Fukumizu, Choon Hui Teo, Le Song, Bernhard Schöumlkopf, and Alex Smola, *A kernel statistical test of independence*, Advances in Neural Information Processing Systems 20, MIT Press, 2008, pp. 585–592.
- [GL05] E. Guerre and P. Lavergne, *Data-driven rate-optimal specification testing in regression models*, Ann. Statist. **33** (2005), no. 2, 840–870.
- [HM93] W. Hardle and E. Mammen, *Comparing nonparametric versus parametric regression fits*, Ann. Statist. **21** (1993), 1926–1947.
- [Ken08] Peter Kennedy, *A guide to econometrics (6th ed.)*, Blackwell, 2008.
- [LN09] Anne Leucht and Michael H. Neumann, *Consistency of general bootstrap methods for degenerate U-type and V-type statistics*, J. Mult. Anal. **100** (2009), no. 8, 1622–1633.
- [Lyo11] Russell Lyons, *Distance covariance in metric spaces*, arXiv preprint arXiv:1106.5758 (2011).
- [Neu09] N. Neumeyer, *Testing independence in nonparametric regression*, J. Mult. Anal. **100** (2009), no. 7, 1551–1566.
- [NVK10] N. Neumeyer and I. Van Keilegom, *Estimating the error distribution in nonparametric multiple regression with applications to model testing*, J. Mult. Anal. **101** (2010), no. 5, 1067–1078.
- [SMQ98] W. Stute, W.G. Manteiga, and M.P. Quindimil, *Bootstrap approximations in model checks for regression*, J. Amer. Statist. Assoc. **93** (1998), no. 441, 141–149.
- [SR09] Gábor J. Székely and Maria L. Rizzo, *Brownian distance covariance*, Ann. Appl. Stat. **3** (2009), no. 4, 1236–1265.
- [SRB07] Gábor J. Székely, Maria L. Rizzo, and Nail K. Bakirov, *Measuring and testing dependence by correlation of distances*, Ann. Statist. **35** (2007), no. 6, 2769–2794.
- [SSGF12] D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu, *Equivalence of distance-based and rkhs-based statistics in hypothesis testing*, arXiv preprint arXiv:1207.6076 (2012).
- [Stu97] W. Stute, *Nonparametric model checks for regression*, Ann. Statist. (1997), 613–641.

- [Xia09] Yingcun Xia, *Model checking in regression via dimension reduction*, Biometrika **96** (2009), 133–148.